doi:10.1111/j.1365-2427.2006.01675.x

Using biological metrics to score and evaluate sites: a nearest-neighbour reference condition approach

SAMANTHA C. BATES PRINS* AND ERIC P. SMITH†

**Department of Mathematics & Statistics, James Madison University, Harrisonburg, VA, U.S.A.*

†Department of Statistics, Virginia Polytechnic Institute & State University, Blacksburg, VA, U.S.A.

SUMMARY

1. Reference (i.e. least or minimally impaired) sites can provide important information about the expected range of biological metrics and can be used to establish impairment or non-impairment of a test site. A problem with using reference data is that biological metrics are affected by natural conditions. We present an approach that uses local information to adjust for natural conditions and a method for statistically evaluating condition at a test site using biological metrics.

2. Our method consists of four steps: selection of a distance measure to find neighbours of a test site, selecting natural variables to measure the distance, selection of the number of neighbours and calculating a scored metric.

3. We use a simulated example to illustrate when the nearest-neighbour approach improves classification of sites as reference or not reference.

4. Using a set of data from the Mid-Atlantic Highlands, we show that the nearest-neighbour method improved on the ability of a regression approach to correctly classify test sites known to be from a non-reference group without affecting the ability to correctly classify test sites known to be from the reference group.

Keywords: Benthic Assessment of Sediment, multimetric, River Invertebrate Prediction and Classification System, standards assessment, stressor–response

Introduction

Reference locations (locations of least or minimal impairment) provide valuable information that can be used to describe reference conditions, to evaluate the impairment of sampled sites (e.g. stream segments of a particular size) and to assess recovery of sites listed as impaired (Burton, Chapman & Smith, 2002; Bailey, Norris & Reynoldson, 2004). Reference condition analysis forms the basis of several environmental assessment methods or models including the River Invertebrate Prediction and Classification System (RIVPACS) (Wright *et al.*, 1984; Clarke *et al.*, 1996), the Benthic Assessment of Sediment (BEAST) (Reynoldson *et al.*, 1995), the Australian River Assessment System (AUSRIVAS) (Nichols *et al.*, 2000; Simpson & Norris, 2000), the test site analysis (Bowman & Somers, 2005; Bowman *et al.*, 2006), the multimetric approach (Barbour, Stribling & Karr, 1995) and the Assessment by Nearest Neighbor Analysis (ANNA) system (Linke *et al.*, 2005) and can be used with biological, chemical or other types of data (Reynoldson, Smith & Bailer, 2002).

An important recognition in the use of reference conditions for evaluation of a test site is that a test site may be different from some reference sites even when this test site is indeed from the set of reference sites (Chessman & Royal, 2004). Hence, quantities (e.g. mean and standard deviation) calculated from the reference metrics may not represent what is expected for the test site. For example, the mean of a

Correspondence: Samantha C. Bates Prins, Department of Mathematics & Statistics, James Madison University, MSC 1911, Harrisonburg, VA 22807, U.S.A.
E-mail: prinssc@jmu.edu

metric at high sites may be different from the mean of that metric at low sites. A test site at high elevation would be best evaluated using high elevation reference sites. Recognising the potential for differences among samples within the reference set, Wright *et al.* (1984) and Chessman (1999) recommended the use of subset methods or adjustment methods to account for potential differences due to natural factors.

In the subset approach, methods such as cluster analysis are used to find groups of similar reference sites. Approaches such as RIVPACS and BEAST use subset methods to group reference sites, resulting in reference data groups that are expected to have similar biological characteristics. Creating groups of similar reference sites should reduce variation in measurements and result in more sensitive biological evaluation of new sites.

In the adjustment method, regression is used to model the relationship between predictor variables unrelated to anthropogenic stress and the biological variables. These regression relationships are developed using only the reference sites but are applied to the test site data. The predicted value of a biological variable associated with a test site then represents the reference mean adjusted for the values of the predictor variables measured at the test site (Chessman, 1999). Deviation from the predicted value is a measure of difference from reference condition. Chessman (1999), for example, uses a regression model to predict the biological metric at a test site from reference sites and then uses the predicted values along with tolerance values to obtain an expected metric value. To evaluate a test site, the observed metric value is then compared with the expected metric value. Yuan & Norton (2003) use regression-based scaling of benthic macroinvertebrate community metrics to evaluate stressor–response relationships in the Mid-Atlantic Highlands. They fit a linear regression model of the response metric of interest on selected predictors using data from a large collection of reference sites. Using this model, they obtain a prediction of the metric at each test site assuming the test site was from the reference group, and an associated estimate of the error in this prediction. The standardised difference between observed and predicted is then used to measure deviation from reference. Standard regression methods are global methods in that the regression model is assumed to hold across the whole set of reference sites. Methods such as BEAST are semi-global in the sense that the properties of the reference set are constant within groups or clusters of reference sites.

An alternative method, used in ANNA, uses a local set of neighbours to a test site to determine the observed and expected taxa richness. We propose a local scaling method for biological metrics or multimetrics, which is similar to ANNA, in that a subset of reference sites is used for scaling. For each biological metric and test site, there will be a set (possibly different for each test site) of reference values that is used to compute the mean and variance for scaling. The scaled biological metric is then used to evaluate the test site. The neighbours are chosen based on proximity to the test site using a set of predictor variables chosen by the user.

The advantage of scaling and evaluating by local rather than global reference conditions is that the reference distribution for a particular test site will be more similar to that site in the values of the selected predictor variables. The decrease in heterogeneity should lead to decisions that are more accurate. A difficulty with using regression models to reduce variability in biological data is that the variance explained (i.e. R^2) is often small when models are fitted to reference site data. Regression approaches require assumptions (linearity and homogeneity) that may or may not be valid on a global (i.e. over the whole range of the predictors) scale. Local scaling and evaluation permits a non-linear global relationship to hold between the predictors and response. Although the method we describe in more detail below is similar to that used in ANNA, ours is local. The ANNA model uses a nearest-neighbour-based method that requires a global assumption of linearity between chosen environmental variables and axis scores from a non-metric multidimensional scaling. ANNA's nearest-neighbour distances are weighted by coefficients from multiple linear regression models applied to observations from all the sites. By focusing on metrics that are simply derived from biological count data (e.g. taxa richness or Simpson's diversity measure), our method does not require a global multivariate step.

All of the above methods or models evaluate a test site by comparing what is observed at a test site to what is expected or predicted at that site based on reference conditions. We refer to this process as scaling of the test site. We assume the degree of directional difference from the numerical condition at reference sites is related to the degree of impairment of a test site. While there may be other explanations for these differences, a large difference is typically viewed as indicative of a potential problem site. There are a variety of ways to calculate this difference. Evaluation of a test site using RIVPACS, AUSRIVAS and ANNA is based on the ratio of the observed and expected (O/E) number of selected macroinvertebrate taxa and small values imply impaired conditions. BEAST declares a test site as potentially impaired if its biological distance to the centre of the reference data is larger than the distance from reference sites to their centre [see Reynoldson *et al.* (1995) for details]. By reference metric scaling, we refer to a general set of methods for adjusting or normalising a metric (or multimetric) to reference conditions. Its purpose may be descriptive (e.g. to produce a cumulative distribution for reference sites) although the most common use of scoring is to identify potentially impaired sites.

Reference scoring usually involves adjustment by expectation and/or variance using a linear approach. When O/E is used to evaluate a test site, the scoring of the observed number of selected taxa is by expectation. The BEAST method uses both expectation and variation in its multivariate distance calculation. Methods used for scoring in statistical applications are often based on some form of standardisation. We distinguish between two such types, based on how uncertainty is accounted for in the formulas and the purpose of the standardisation. Standardisation for descriptive purposes involves subtracting the reference mean and dividing by the reference standard deviation to produce the scored metric. This is commonly used to assess where an observation lies relative to other observations and to describe the biological population after adjusting for reference conditions. Thus, the scored metric is adjusted for expectation then scaled by the standard deviation calculated from the reference sites using

$$\text{scored metric} = \frac{\text{test metric} - \text{reference mean}}{\text{reference standard deviation}}. \quad (1)$$

It is important to note that the information from the test site is not included in the calculation of either the expectation (i.e. reference mean) or the standard deviation. A scored metric at a test site can be interpreted as the number of standard deviations above or below what is expected from the information in the reference sites and hence may be used to describe the condition of a test site relative to the reference set. An example using this approach to develop a stream health index is given in Chiu & Guttorp (2006).

An alternative to scoring for description views the scoring as a prediction problem and uses a prediction variance rather than an estimated variance of reference conditions, i.e.

$$\text{scored predicted metric} = \frac{\text{metric} - \text{reference predicted metric}}{\text{reference prediction standard deviation}}. \quad (2)$$

Equation (1) is most useful for descriptive purposes, while (2) is useful for testing hypotheses such as if a site has biological conditions that are different from reference conditions. An important difference between the two methods is that prediction based scoring views the test site as a new observation and the variance that is calculated includes uncertainty associated with the predicted value. This approach is common in methods such as regression analysis to evaluate if new observations are consistent with the regression model. See Montgomery, Peck & Vining (2006) for a discussion of prediction versus description in the regression setting. Kilgour, Somers & Matthews (1998) suggest using the square of the scored predicted metric as a way of evaluating if a test site is in the normal range of reference conditions.

We refer to the situation where all available reference sites are used in eqns (1) and (2) as the 'null' model approach. These null model equations are given in Table 1 (formulas (4) and (5)).

In this paper, we describe a method that scores the value of a metric at a test site using the average response and standard deviation of the responses observed at reference sites that are the nearest-neighbours of the test site. Nearest-neighbours are chosen based on the proximity of the test site to reference sites using a set of predictors chosen by the user. Neighbour distances can be based on a combination of continuous and categorical predictors. Using a simple simulated data set, we highlight potential advantages of the method relative to the regression and 'null' approaches. We also evaluate the nearest-neighbour method on a data set from the Mid-Atlantic Highlands. We conclude with discussion and areas for future work.

Table 1 Formulas for descriptive and predictive scoring of metrics

Model	Type of scoring	Equation	Test Distribution
Null or mean	Descriptive	$y_{i,d}^{M} = \frac{y_i - \bar{y}_R}{s_R}$ (4)	t_{N-1}
	Predictive	$y_{i,p}^{M} = \frac{y_i - \bar{y}_R}{s_R\sqrt{1+1/N}}$ (5)	t_{N-1}
Regression	Descriptive	$y_{i,d}^{Reg} = \frac{y_i - \hat{y}_{i,reg}}{s_{reg}}$ (6)	t_{N-2}
	Predictive (single predictor)	$y_{i,p}^{Reg} = \frac{y_i - \hat{y}_{i,reg}}{s_{reg}\sqrt{1+\frac{1}{N}+\frac{(x_i-\bar{x})^2}{\sum_{j=1}^{N}(x_j-\bar{x})^2}}}$ (7)	t_{N-2}
Nearest-neighbour	Descriptive	$y_{i,d}^{NN} = \frac{y_i - \bar{y}_{i,k}}{s_{i,k}}$ (8)	t_{k-1}
	Predictive	$y_{i,p}^{NN} = \frac{y_i - \bar{y}_{i,k}}{s_{i,k}\sqrt{1+1/k}}$ (9)	t_{k-1}

In the equations, y_i represents the value of the metric at site *i*. The second subscript on *y* refers to either descriptive (*d*) or prediction (*p*). Superscripts *M*, *Reg*, *NN* refer to Null/Mean, regression and nearest-neighbours respectively. In the equations, *R* refers to reference, *N* is the total number of reference sites, x_i is the value of the predictor at site *i*, and *k* is the number of nearest-neighbours. The mean of the metrics is denoted using $\bar{y}$ and the mean of the predictor is $\bar{x}$. The standard deviation is denoted using *s*. Critical values are obtained from *t*-distributions with different degrees of freedom.

Methods

In our methodology, the relevant data for scoring and testing are one or more biological metrics and potential predictors of the metrics from *N* reference sites that are to be used to describe reference sites. Examples of metrics include taxa richness or Ephemeroptera, Plecoptera, Trichoptera richness (EPT richness) as well as multimetrics. The predictors are used to explain and hence reduce variation in reference conditions, not to explain why a test site might be considered impaired. Examples of reference predictors might include altitude, catchment area, time of sampling (e.g. spring or autumn) and whether the biological sample comes from a pool or a riffle.

Our approach is based on the evaluation of individual metrics or a multimetric index. Thus, at selected test site *i* we observe a response metric y_i and *p* predictors, $x_{i1}, x_{i2},\ldots,x_{ip}$. The *k* nearest-neighbours of site *i* are those *k* reference sites that are the 'closest' to ($x_{i1}, x_{i2},\ldots,x_{ip}$) according to a chosen distance measure. These *k* nearest-neighbours of site *i* collectively yield *k* values of the metric that form the neighbourhood reference set for site *i*. The objective then is to score the test site metric, y_i, using this neighbourhood reference set.

Our approach involves a four-step process:

1. Choose a distance measure to determine proximity of reference sites to the test sites.
2. Select variables from potential reference predictors.
3. Select *k*, the number of nearest-neighbours to use.
4. For each test site, calculate the scored metric using the mean and standard deviation of the metric at the reference neighbour sites.

Details on the steps are provided below. In applications, we often will not know the best predictors or optimal *k* so we describe an approach for selection of the predictors and *k* that combines steps 2 and 3.

The first step is the selection of a distance measure. We use the heterogeneous Euclidean overlap metric (HEOM) of Wilson & Martinez (1997) to measure distance between sites. To compute the distance, the predictors are first grouped into continuous and categorical variables. The HOEM distance, dist (*i*, *j*), between a test site *i* and a reference site *j* is given by

$$\mathrm{dist}(i,j) = \sum_{c} \frac{|x_{i,c} - x_{j,c}|}{\mathrm{range}_c} + \sum_{g} \mathrm{I}[x_{i,g} \neq x_{j,g}] \qquad (3)$$

where the first of the added terms is a sum over the continuous predictors and the second is a sum over the categorical predictors, range_c is the range of the *c*th continuous predictor in the reference dataset, and I[ ] is equal to 1 if sites *i* and *j* have a different value of the *g*th categorical predictor, and is equal to 0 otherwise.

The first term in the HEOM distance represents the total contribution of the continuous predictors. Although standard deviation could be used in the denominator (Smith *et al.*, 2003; Yuan & Norton, 2003), we used range to normalise the contribution of an individual predictor to the overall distance to be in a 0–1 range which is commensurate with the contribution of the categorical predictors. Transformation of the predictors is often useful when using the range to avoid effects from skewness of observations. One advantage of HEOM is that it allows both continuous and categorical predictors to contribute to the distance between any two sites. HEOM can also be adjusted for ordinal data. Although the name of this distance measure suggests the use of Euclidean distance for continuous predictors, Manhattan distance is used for computational simplicity in Wilson & Martinez (1997) and here. Other distance measures might also be acceptable.

The second and third steps involve the selection of reference predictors and number of neighbours. The choice of predictors and k is important and leads to better metric prediction and variance estimation. In practice, it may be possible to select these variables using expert knowledge. However in some cases detailed information about individual sites is not available and we may want to choose the predictors and k using an objective statistical approach. Such an approach follows: select a range of values for k (from 1 to N). For each value of k selected, run a leave-one-out subset selection method to identify the subset of the predictors that minimise the mean squared error (MSE) of the predicted metric at the reference sites. That is, for each reference site in turn, and for each candidate predictor, apply the nearest-neighbour scoring method using that predictor in the distance measure and the remaining N–1 reference sites to characterise reference conditions. Determine the MSE of the predicted metrics at the N reference sites and select the single predictor that minimises the MSE. Continue to add predictors until the MSE is not reduced. The MSE measures the closeness of the predicted metric value to the true metric value and a small MSE would suggest low prediction bias and low uncertainty. The process results in a two-way table of MSEs for predictors and values of k. The user should then choose the value of k and the associated subset of predictors for which the MSE of the predicted metrics on the reference sites is at its minimum. The number of neighbours need not be the same for every metric of interest. Similarly, when the assessment involves multiple metrics, a different subset of predictors may be selected for different biological metrics or the same set used for all metrics.

Note that using a categorical predictor in the variable selection procedure may be complicated when k is small since there may be multiple sets of sites with the same level of the categorical predictor. We therefore only consider adding the categorical predictor after a continuous predictor has been included. Following selection of variables and k, distances between sites are calculated.

Finally, in the fourth step, to obtain the scored value of the metric at a test site i, we calculate the HEOM distance between the chosen predictors at test site i and every individual reference site. Next, find the k nearest-neighbours and compute the mean and standard deviation. Then score the metric using either eqn (1) or eqn (2) with the nearest-neighbour statistics (see Table 1, formulas (8) and (9)). The size and direction of the scored metric may indicate the degree of impairment or concern for a test site. If we call metrics that increase with increased stress positive metrics then a scored positive metric value above the critical value from a t-distribution with $k - 1$ degrees of freedom and upper tail probability α (hereafter denoted by $t_{k-1}(\alpha)$) would signal possible impairment. Similarly, for metrics that are expected to decrease with increasing stress, a scored metric value below the criterion of $t_{k-1}(1 - \alpha)$ would signal possible impairment. (In most situations, these would result in an approximate criterion of 'above 2' for positive metrics and 'below −2' for negative metrics.) This rule is somewhat similar to a percentile rule (see e.g. Clarke *et al.*, 1996) and corresponds to the 'reference interval' approach used in medical diagnostics (Altman, 1991).

The approach in eqn (2) suggests a more formal statistical evaluation of the test site and divides by the standard deviation of prediction. To formally test for impairment we would score the metric and compare its value with the critical value from a t-distribution with $k - 1$ degrees of freedom and upper tail probability α or $1 - \alpha$ depending on whether the metric is positive or negative.

In the comparisons given below, the metrics are scored using the nearest-neighbour, regression and null model approaches. For the nearest-neighbour method, the calculations are based on predictive scoring of the metrics using the mean of the k nearest-neighbours and the corresponding standard error in eqn (9). For the regression approach, the predictive scored metrics at test sites are found using the predicted value from a regression model and the corresponding standard error using eqn (7). The regression-based equations are slightly different from that used by Yuan & Norton (2003) who divided by the standard deviation of the regression residuals. The table gives the t-distribution for testing with a single predictor; testing for the general case would use a critical value for a t-distribution with $N - r$ degrees of freedom where r is the number of parameters in the regression. Null model calculations are based on eqn (5) and use all reference sites.

Methods of Evaluation

We evaluated the nearest-neighbour method for different k and compared it with the regression and null methods using two criteria. If useful, the nearest-neighbour method should produce near zero values of the scored metrics when applied to the reference set (i.e. a reference site should be similar to its nearest-neighbour reference sites). The method should also produce large values of the scored metrics for stressed test sites. Our comparison for the reference sites is based on treating each reference site as a potential test site and then classifying the site. The scored metric for the held-out reference sites should produce better estimates of how well the method would work on new reference sites.

The first criterion evaluates the prediction ability of the methods on the reference site data. For both the nearest-neighbour and regression approaches, we calculated the MSE of the predicted metrics, obtained for the held-out reference sites. For regression scoring, bias is based on the prediction from a regression equation and MSE is given by the average prediction variance. If overall prediction at reference sites is better, the MSE at these sites should be lower.

The second evaluation procedure is relevant when classification of sites as 'not impaired' or 'impaired' is of particular interest. We illustrate assessments based on standardised metrics by classifying a positive scored metric at a test site as indicating 'impairment' if its value is greater than $t_{\mathrm{d.f.}}(\alpha = 0.05)$ where the degrees of freedom (d.f.) are as given in Table 1. Similarly, the score for a negative metric denotes impairment if it lies below $t_{\mathrm{d.f.}}(1 - \alpha = 0.95)$. To illustrate the ability of the regression and nearest-neighbour approaches to distinguish between reference and potentially impaired sites, we use a set of data where low pH sites are treated as the potentially impaired sites. If successful, the nearest-neighbour method should have better classification than the regression method.

Data

To illustrate a situation where the nearest-neighbour method is superior to a regression scoring approach, we simulated a reference data set of 88 sites with varying mean and variance. The predictor variable, x, was generated from a uniform distribution with approximate range 1–5. We generated 88 values of the metric from a normal distribution with mean that assumed an underlying quadratic relationship with the predictor (i.e. $y = 0.25x^2 - x + 6$) and standard deviation (SD) that increased from 0.3 to 1.0. (The first 22 values had SD of 0.3, the next 22 had SD of 0.4, the next 22 had SD of 0.5, the next 11 had SD of 0.75 and the final 11 values had SD of 1).

To illustrate the method further, we used the Environmental Monitoring and Assessment Program (EMAP) data collected in the Mid-Atlantic Highlands (MAHA) [see Klemm *et al.* (2002) or Yuan & Norton (2003) for more extensive discussion of the data]. Yuan & Norton (2003) investigated the sensitivities of six biological metrics to anthropogenic stressors in the Mid-Atlantic Highlands region and we study the same metrics. Three of these metrics are considered positive and three negative in relation to increased stress. The proportional abundance of tolerant taxa of aquatic macroinvertebrates (abbreviated to TOLRPIND), tolerant taxa richness (TOLRRICH) and the proportional abundance of the three most abundant taxa (DOM3PIND) are positive metrics while Ephemeroptera richness (EPHERICH), Plecoptera richness (PLECRICH) and total taxa richness (TOTLRICH) are negative metrics.

The sites in this dataset were classified as reference or non-reference based on values of gran acid neutralising capacity (ANC), chloride (CL-), sulphate (SO4=), total nitrogen (NTL), total phosphorus (PTL) and total rapid bioassessment protocol habitat score (RBP). Yuan & Norton (2003) give a detailed description of the criteria for a reference site developed by Waite *et al.* (2000). Using the criteria for reference sites given in their paper, we determined that 87 of the 503 sites in our Mid-Atlantic Highlands dataset were reference sites. Although we obtained a slightly different number of reference sites, the characteristics of the sites were similar to those reported in Table 2 of Yuan & Norton (2003). Table 2 gives a summary of the data.

Potential predictor variables included the continuous variables log-transformed catchment area (AREA), latitude (LAT), longitude (LON) and RBP, and a categorical variable representing the Level III ecoregion for defining neighbours of a test site. We used the variable selection method described above to choose the best set of predictors and value of k for the nearest-neighbour approach. For the regression approach, we used the variable selection approach outlined in Yuan & Norton (2003) for selecting the best subset of predictors in a regression approach.

All analyses were done in R Version 2.0.1 (Ihaka & Gentleman, 1996).

Results

The simulated reference data are displayed in Fig. 1 along with fitted values for the three methods and associated intervals assuming the metric is positive. The solid lines correspond to the fitted values while the dashed lines indicate the estimated upper error bar or boundary for reference values. At a specific value of the predictor variable (x-axis), a test metric with value above the boundary defined by the interval would be classified as non-reference. The regression fit is significant ($R^2 = 0.57$).

Using the error bars to define the boundary for a decision rule, we obtain four regions (for each approach we have a region where a test site would be declared as 'impaired' or 'not impaired'). What results from the display is that the null model approach has a constant width band and the regression approach has a roughly constant width band while the width of the band for the nearest-neighbour method varies. In general, the nearest-neighbour error band is smaller although in some cases it is larger than the regression band. The large empty area below the boundary for the null model (left side of graph) indicates an area where the null model method would declare a site as not impaired but the other models would not. The resulting higher misclassification rate for this area is due to the null model being incorrect in this region (i.e. the trend in the data is not part of the model). There is a large area (labelled B) where the regression method would declare a site as not impaired while the nearest-neighbour method would declare impairment. Thus, we expect the regression method to have a small correct classification rate. In some cases (C), the nearest-neighbour method classifies the site as not impaired when the regression method classifies the site as impaired.

Results summarising the leave-one-out MSE for the MAHA reference data are presented in Fig. 2. When

Table 2 Numerical summaries of all 503 sites in the Mid-Atlantic Highlands dataset and the 87 sites meeting the reference conditions

	Minimum	First quartile	Median	Mean	Third quartile	Maximum
Total rapid bioassessment protocol habitat score	−235	−198	−174	−170.5	−149.0	−24
Reference	−235	−212	−203	−202.4	−187.5	−181
Catchment area	0.65	2.36	3.03	2.98	3.65	4.77
Reference	1.54	2.34	3.16	3.08	3.78	4.48
Ephemeroptera richness	0	2.5	6	5.8	9.0	18
Reference	0	6.0	9	8.7	11.0	17
Plecoptera richness	0	2.0	3	3.5	5.0	10
Reference	1	3.0	5	5.0	6.5	10
Tolerant taxa richness	0	2.0	3	3.7	5.0	15
Reference	0	1.0	2	2.4	4.0	7
Proportional abundance of tolerant taxa of aquatic macroinvertebrates	0.00	0.10	0.19	0.23	0.31	1.49
Reference	0.00	0.07	0.12	0.13	0.18	0.42
Proportional abundance of the three most abundant taxa	0.43	0.65	0.75	0.78	0.88	1.57
Reference	0.45	0.58	0.67	0.70	0.81	1.05
Total taxa richness	1	28	37	36.3	44.5	72
Reference	7	35	43	42.7	51.0	70

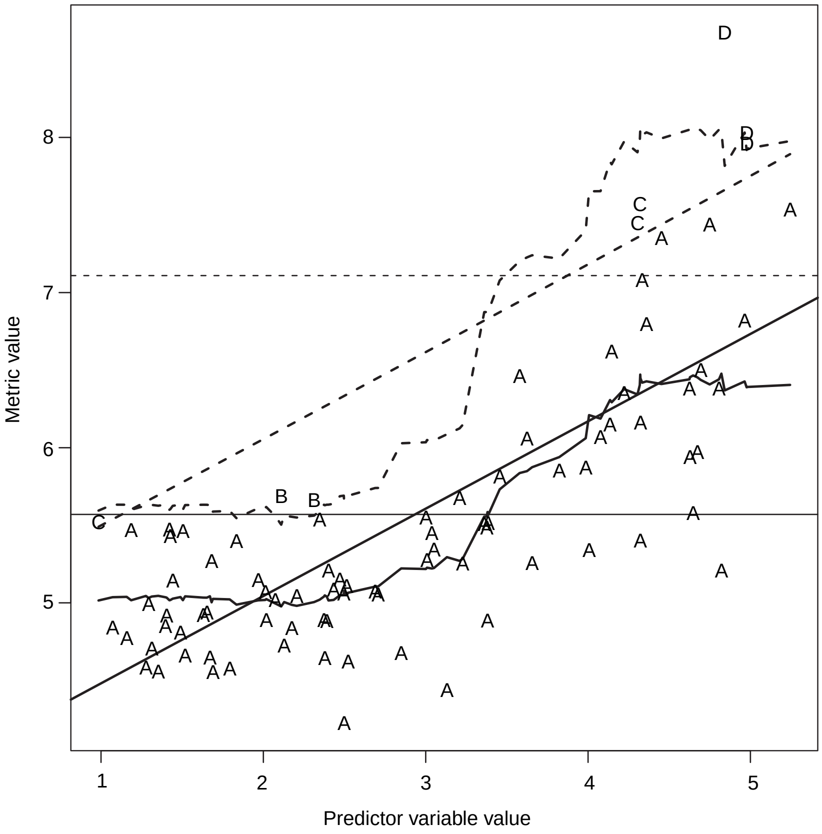

Fig. 1 Regression based prediction (positive slope solid line) and nearest-neighbour prediction with $k = 32$ (wiggly solid line) for the 88 simulated reference sites. Upper confidence limits for the regression (positive slope dashed line) and nearest-neighbour with $k = 32$ (wiggly dashed line) methods are shown. Predictions and upper confidence limits for the regression and nearest-neighbour approaches were obtained using leave-one-out analysis with $\alpha = 0.05$. Points are labelled 'A' if both methods would classify the point as 'not impaired', 'B' if only the regression would classify as 'not impaired', 'C' if only nearest-neighbour would classify as 'not impaired', and 'D' if both would classify as 'impaired'. The horizontal lines represent the prediction (solid) and upper confidence limit with $\alpha = 0.05$ (dashed) for the null approach to scoring.

between approximately 20% and 50% of the reference sites are used, the reference site predicted values for all metrics have lower MSE than those when k represents 100% of the reference sites suggesting that using a subset rather than all of the reference sites is best. The selected neighbourhood sizes are then $k = 15$ for EPHERICH and DOM3PIND, $k = 14$ for PLECRICH, $k = 21$ for TOTLRICH and $k = 20$ for the remaining metrics. AREA was selected for scoring all six metrics, along with LAT for EPHERICH, TOTLRICH and DOM3PIND, and LON for DOM3-PIND. The regression variable selection procedure, applied to the 87 reference sites selected log catchment area as the only predictor in regression models for each of the six metrics.

For four of the six metrics, the nearest-neighbours approach results in a smaller MSE relative to the regression approach for some interval of k values (Fig. 2). For EPHERICH and TOTLRICH, the regression MSE is lower for all k, but is close to the minimum nearest-neighbour MSE (Fig. 2).

We found that when ecoregion is considered as a potential predictor the MSE increased for all metrics, unless k was large. For example, the MSE for TOLRPIND when k was 20 with ecoregion considered as a predictor was 0.00849 and this decreased to 0.00609 with ecoregion omitted. When $k = 77$ the MSE for TOLRPIND was 0.00850 with ecoregion considered and 0.00845 with ecoregion omitted. Use of ecoregion did not improve the MSE so it was not included in the nearest-neighbour model.

Table 3 displays results describing the classification of reference sites. We note that all three methods give relatively high correct classification rates.

To compare the nearest-neighbour and regression methods ability to correctly classify potentially impaired and impaired sites, we considered the Ephemeroptera richness metric and a non-reference group defined using pH < 6.0, consisting of 50 sites. Using the predictive scoring approach with critical value based on $\alpha = 0.025$ as the cutoff for impairment, the nearest-neighbour approach using AREA and LAT and an optimal k of 15 neighbours classified 56% of these sites (28 sites) as potentially impaired, compared with 44% (22 sites) for the null model and 34% (17 sites; $R^2 = 0.39$) using regres-

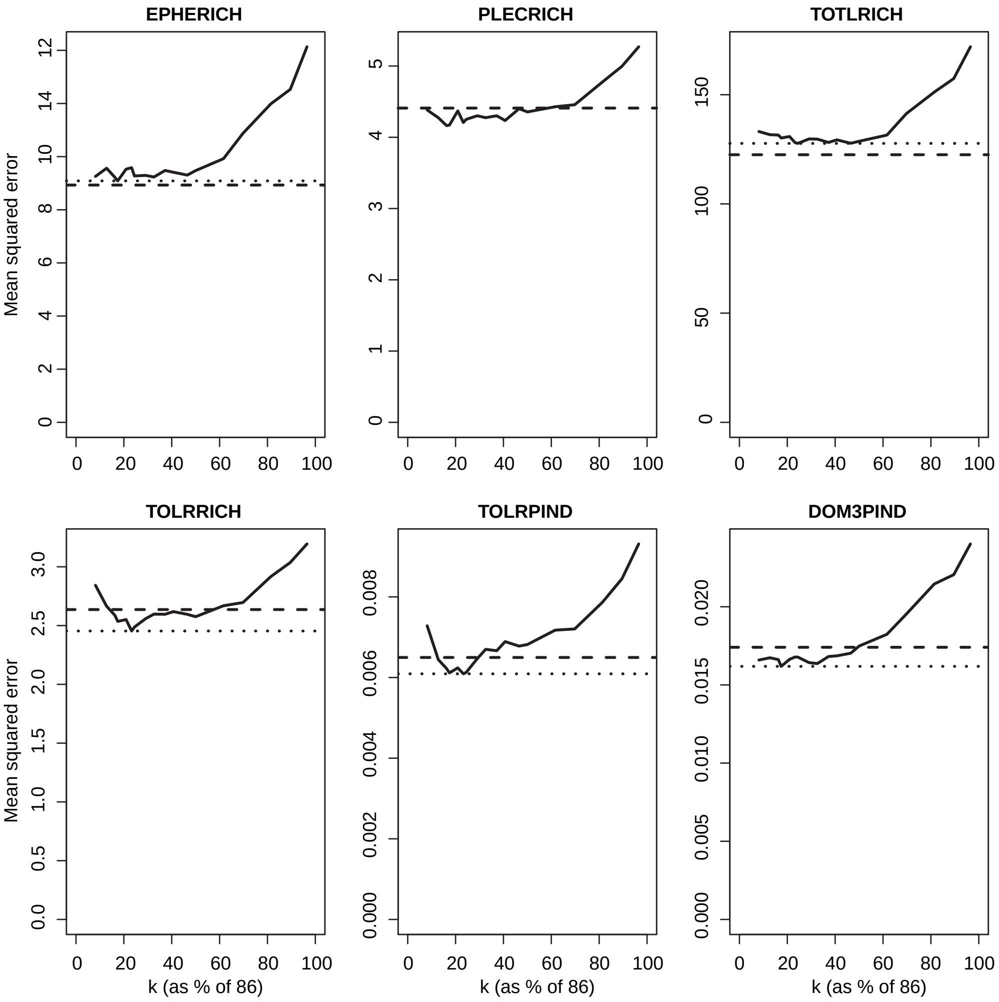

Fig. 2 Mean squared error (MSE) of the six predicted (unscored) metrics at the 87 reference sites as a function of the number of nearest-neighbours for the optimal set of predictors. The solid line represents prediction with *k* nearest-neighbours, and the dashed line prediction with the regression approach (this is independent of *k*). The dotted line represents prediction of Ephemeroptera richness (EPHERICH) and proportional abundance of the three most abundant taxa (DOM3PIND) using *k*=15 nearest-neighbours, Plecoptera richness (PLECRICH) using *k*=14, total taxa richness (TOTLRICH) using *k*=21, and Proportional abundance of tolerant taxa of aquatic macroinvertebrates (TOLRPIND) and tolerant taxa richness (TOLRRICH) with *k*=20 nearest-neighbours. Values of *k* considered represent 8%, 13%, 16%, 17%, 21%, 23%, 24%, 29%, 33%, 37%, 41%, 47%, 50%, 62%, 70%, 81%, 90% and 100% of the 86 reference sites not left out.

Table 3 Percentages of reference sites classified as 'not impaired' by the regression and nearest-neighbour methods with critical values in Table 1 determined with $\alpha = 0.05$

	Negative metrics			Positive metrics		
% of reference sites	EPHERICH	PLECRICH	TOTLRICH	TOLRRICH	TORLPIND	DOM3PIND
Regression method	95	97	95	94	95	94
Nearest-neighbour	95	93	95	95	94	93
Null model	95	93	95	97	91	93

Ephemeroptera richness (EPHERICH) and proportional abundance of the three most abundant taxa (DOM3PIND) were scored using $k = 15$, Plecoptera richness (PLECRICH) using $k = 14$, and total taxa richness (TOTLRICH) using $k = 21$. Proportional abundance of tolerant taxa of aquatic macroinvertebrates (TOLRPIND) and tolerant taxa richness (TOLRRICH) were scored using $k = 20$ nearest-neighbours. The null model uses $k = 86$.

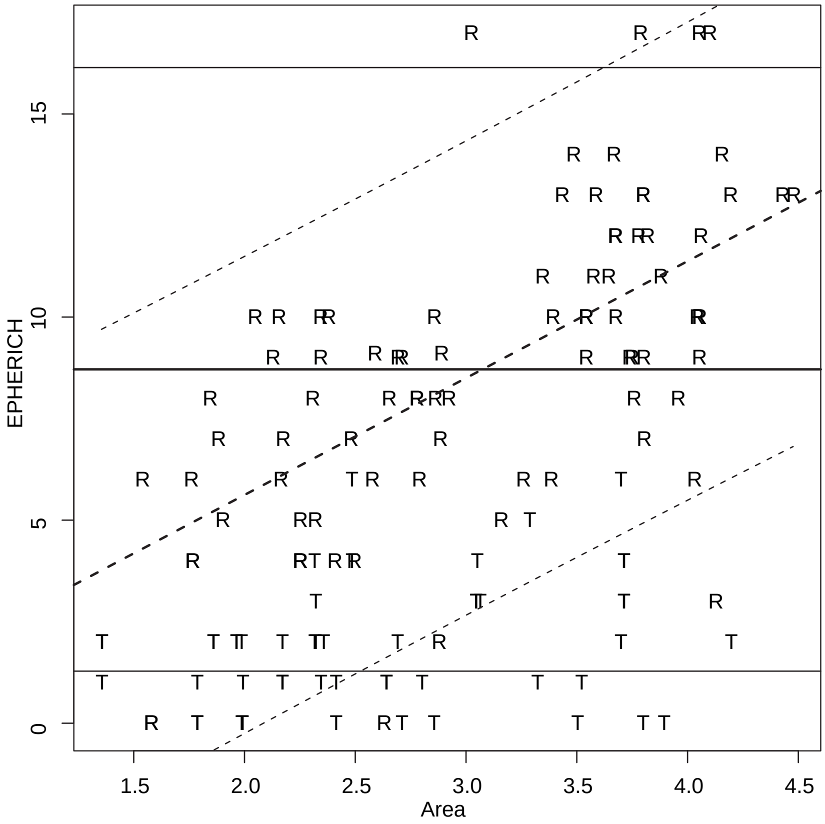

Fig. 3 Plot of the reference (R) and test (T) set data along with the fitted regression model and prediction based standard error bars using catchment area (AREA) as the predictor (dashed lines) and probability $\alpha = 0.05$. The solid lines represent the fitted null model prediction with standard error bounds based on $\alpha = 0.05$.

sion. Fig. 3 plots the reference and test data along with the fitted regression and null model lines with standard error bars. If instead, $\alpha = 0.05$ is used to test impairment the regression method will result in more rejections and a classification rate that is closer to the nearest-neighbour method (using a 5% test, the classification rate is 70% for the nearest-neighbour method compared with 58% for regression; using a 2.5% test, the classification rate is 56% and 34% respectively).

There are three reference sites with zero values of the EPHERICH metric that potentially affect both the regression and nearest-neighbour approaches. They provide some caution that reference criterion should not be based solely on chemistry. If these three sites are removed from the reference data set and both approaches rerun, we find that the nearest-neighbour approach with AREA and LAT and $k = 15$ nearest-neighbours classifies 76% of the 50 test sites as potentially impaired, 18% more than the regression approach ($R^2 = 0.35$) and 4% more than the null model, using an impairment criterion based on $\alpha = 0.025$.

Discussion

We have proposed an approach to adjust the observed value of a particular biological metric at a test site using data from the *k* reference sites that are closest in selected predictors. Results using a set of data from the Mid-Atlantic Highlands indicate that this nearest-neighbour method performed comparably with the regression approach in correctly classifying reference sites. Classification of a test set with low pH indicated a greater number of sites were classified as potentially impaired using the nearest-neighbour approach. Simulated data also suggest that the nearest-neighbour approach provides reference distributions for a test site that are closer to the proper reference distribution for that site in situations with heterogeneous variance and non-linearity. Although the approach may be viewed as a new method for evaluation of biological monitoring data, our view is that the method is a statistical refinement of the ANNA method. While the data that is collected are taxa counts, the basis for our analysis is a biological metric. Methods such as ANNA (Linke *et al.*, 2005), RIVPACS (Wright *et al.*, 1984) and BEAST (Reynoldson *et al.*, 1995) take a more multivariate view, either to summarise the information or to measure distance. Given the metric we can then make use of local methods for accounting for natural variables and making inferences about status of a site. The local approach makes the method different from other multimetric approaches.

Our method requires selection of an appropriate distance measure, predictor variables and the number of neighbours. While there are many distance measures that may be used, we chose one that uses absolute difference divided by range and allows for categorical as well as continuous predictors. The distance measure is divided by range to give roughly equal weight to each predictor. This can increase the effect of outliers on the choice of nearest-neighbours, but we feel the benefit of including categorical predictors, when available, outweighs this potential problem, as outliers are not as likely to occur in the reference set as in the test set, especially when predictors are transformed.

An important step in the method is the selection of variables to use to find neighbours and the number of neighbours to use. We assumed that the user could provide a list of variables related to the biological measurements. As some of these variables may not be relevant, a method for selection of variables was suggested to determine important predictors. We chose to use an MSE approach that made use of only the reference sites and that should result in good prediction of the metric responses. Application of the variable selection method to different biological metrics may lead to a different set of environmental variables for each metric. In cases where a set of sites with known impairment are available, an alternative method is to use the biological information and a classification criteria for selection of variables.

As discussed in the Methods section we did not favour scoring with predictions from linear regression models. Although a regression approach might be advantageous when there are strong relationships we have typically found weak regression relationships for reference data. Yuan & Norton (2003) noted that this was the case in the Mid-Atlantic Highlands dataset as the linear regression models used to obtain their scored response values explained between 8% and 31% of the variation in that response. Further, the regression approach assumes a 'global' linear relationship between the metric and the selected natural variables or stressors over all reference sites. Although the ANNA (Linke *et al.*, 2005) approach is based on nearest-neighbours, the distances are based on predicted axis scores from a global regression. However, if a single metric is used (rather than non-metric multidimensional scaling scores) and there is a single predictor, the distances will be the same. Our nearest-neighbour approach assumes neither a linear nor a global relationship. If one chooses to score with predictions based on linear regression models fitted to the k nearest-neighbour reference sites, one would have a method analogous to local linear regression.

As indicated by the simulated reference data in Fig. 1, the mean or null approach is clearly inappropriate if there is trend in the data. Although the regression fit was significant ($R^2 = 0.57$), the non-linearity and heterogeneity of the data indicated the fit was not good. The nearest-neighbour method does better at tracking the non-linear relationship and the heterogeneity of variance in the data. Thus, the general shape of the nearest-neighbour fit is non-linear, reflecting the true model and the error band is wider for higher values of the predictor, reflecting the increased heterogeneity. While contrived, this example demonstrates that if there are local patterns in the data, then the nearest-neighbour method will use the pattern to improve the decision.

The low prediction of non-reference test sites based on the Ephemeroptera richness metric and pH < 6.0 by the regression method was a result of patterns in the data that are better described by the nearest-neighbour method than the regression approach. Notice in Fig. 3 that a relatively high percentage of test cases fall inside the error bars for the regression or null model when AREA is small. The reason for the higher classification rate for the nearest-neighbour method lies in differences in variance estimation. The regression approach assumes constant variance and the estimated variance is roughly the same across the range of AREA. Although the regression approach may provide a good estimate of EPHERICH for sites with small values of AREA, it does not provide as good an estimate of the variation. Hence, the predictive scored metric for the regression approach is not as extreme on this range of AREA. The nearest-neighbour method measures the variation relative to a smaller set of sites and hence results in a smaller variance estimate and larger scored metrics.

As pointed out by Bailey *et al.* (2004), error rates are important when the methods are used in the decision-making process and for this data, the error rates will be superior for the nearest-neighbour method. Although the nominal Type I error rate for all three prediction methods is the prechosen 'alpha' level, the actual rate depends on the correctness of the assumptions, the fit of the model and quality of the variance estimate. For example, if the regression model is correct, then the null model which uses a single reference distribution whose mean does not depend on the predictor is going to have, in general, a higher Type I error rate in some regions and a lower error rate in other regions, as the variance is not adjusted for the predictor. Thus, for example, the Type I error for the null model will be greater than alpha when the predictor is >4 (in Fig. 1, the reference sites labelled 'A' that are above the null interval are rejected using the null method but not the regression method). Similarly, the Type I error rate for the nearest-neighbour method will be more accurate than the null or regression models when those methods do not adequately describe the data, especially when variance changes. In Fig. 1, test sites with metrics in the region labelled 'C' would be rejected using the null or regression method but not the nearest-neighbour method.

When the nearest-neighbour method is applied and the regression or null models are correct there should be only a slight loss in the nearest-neighbour Type I error rate provided the range in the predictor is not great and k is not small. Basically, the fit provided by the nearest-neighbour approach is a local mean which fits adequately but uses a smaller sample size. If k is small (around 5) the t-statistic for the nearest-neighbour method will be noticeably larger than that for either the null or regression methods and will result in a small Type I error rate. If k is larger, the differences in error rates should be minor.

The Type II error for the nearest-neighbour model will generally be better when the regression or null model is incorrect. For example, in Fig. 1 when the predictor is around 2.5, the points in the region labelled 'B' would not be rejected by the regression method but would be with the nearest-neighbour method. This region is larger for the null model which will have low power for testing sites with predictors below 3. When the null model is correct, the nearest-neighbour method will have slightly lower power than the other models. While exact comparisons are possible, the difference depends on the values of the predictors, strength of the regression model as well as the number of reference sites available and the number of neighbours used. In general, if the null model is correct, it will have the greater power, as the critical value is $t_{N-1}(\alpha)$ and this will be smaller than the regression critical value of $t_{N-2}(\alpha)$ and the nearest-neighbour critical value of $t_{k-1}(\alpha)$. Some indication of the actual error rates that might occur in practice may be obtained through leave-one-out methods with the MAHA data.

A natural extension to the proposed approach [similar to that of Chessman (1999)] would be to weight the chosen k nearest-neighbours by the inverse of their HEOM distance from the test site currently of interest. This involves simple calculations of a weighted mean and standard deviation and does not greatly increase the computational complexity of the method. Such a weighting scheme is used in ANNA (Linke *et al.*, 2005). Other extensions of the method that are easy to implement are to use a multivariate nearest-neighbours approach with multiple metrics and to extend the formulas to situations with multiple measurements at the test site (rather than a single measurement). Another variation of the method is to base biocriteria and testing on percentiles or equivalence (Kilgour *et al.*, 1998; Smith *et al.*, 2003; Bowman & Somers, 2005). These methods use an equation similar to eqn (1) but use a non-central t-distribution rather than a central t-distribution for evaluation. Critical values for tests using a prediction approach tend to be smaller than those for a percentile approach so the prediction approach will be more powerful for detecting biological difference [see Fig. 3.3 in Hahn & Meeker (1991)]. An alternative method of evaluation would examine the relationship of the scored metric values at test sites to known anthropogenic stressor gradients (Klemm *et al.*, 2003). This recognises that the test sites in this data set range in quality from truly degraded to nearly reference quality. We did not apply such an approach in our example due to the lack of a moderate to strong pH gradient. The EMAP data are spatially balanced and hence are not focused on finding gradients, especially over smaller spatial scales.

Acknowledgments

This work was supported by the Environmental Protection Agency (EPA) through STAR grant number R83136801-0. Although the research described in the article has been funded wholly or in part by the U.S. EPA STAR programs, it has not been subjected to any EPA review and therefore does not necessarily reflect the views of the Agency, and no official endorsement should be inferred. We greatly appreciate the efforts of Environmental Monitoring and Assessment Program (EMAP) researchers in collecting and providing this data.

References

Altman D.G. (1991) *Practical Statistics for Medical Research*. Chapman and Hall, London.

Bailey R.C., Norris R.H. & Reynoldson T.B. (Eds) (2004) *Bioassessment of Freshwater Ecosystems: Using the Reference Condition Approach*. Kluwer Academic Publishers, Dordrecht.

Barbour M.T., Stribling J.B. & Karr J.R. (1995) Multimetric approach for establishing biocriteria and measuring biological condition. In: *Biological Assessment and Criteria: Tools for Water Resource Planning and Decision Making* (Eds W.S. Davis & T.P. Simon), pp. 63–77. Lewis Publishers, FL, U.S.A.

Bowman M.F. & Somers K.M. (2005) Considerations when using the reference condition approach for bioassessment of freshwater ecosystems. *Water Quality Research Journal of Canada*, **40**, 347–360.

Bowman M.F., Somers K.M., Reid R.A. & Scott L.D. (2006) Temporal response of stream macroinvertebrates to the synergistic effects of anthropogenic acidification and natural drought events. *Freshwater Biology*, **51**, 768–782.

Burton G.A., Chapman P.M. & Smith E.P. (2002) A review and critique of weight-of-evidence approaches for assessing ecosystem impairment. *Human and Ecological Risk Assessment*, **8**, 1657–1673.

Chessman B.C. (1999) Predicting the macroinvertebrate faunas of rivers by multiple regression of biological and environmental differences. *Freshwater Biology*, **41**, 747–757.

Chessman B.C. & Royal M.J. (2004) Bioassessment without reference sites: use of environmental filters to predict natural assemblages of river macroinvertebrates. *Journal of the North American Benthological Society*, **23**, 599–615.

Chiu G. & Guttorp P. (2006) Stream health index for the Puget Sound lowland. *Environmetrics*, **17**, 285–307.

Clarke R.T., Furse M.T., Wright J.F. & Moss D. (1996) Derivation of a biological quality index for river sites: comparison of the observed with the expected fauna. *Journal of Applied Statistics*, **23**, 311–332.

Hahn G.J. & Meeker W.Q. (1991) *Statistical Intervals: A Guide for Practitioners*. John Wiley and Sons, New York.

Ihaka R. & Gentleman R. (1996) R: a language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, **5**, 299–314.

Kilgour B.W., Somers K.M. & Matthews D.E. (1998) Using the normal range as a criterion for ecological significance in environmental monitoring and assessment. *Ecoscience*, **5**, 542–550.

Klemm D.J., Blocksom K.A., Thoeny W.T., Fulk F.A., Herlihy A.T., Kaufmann P.R. & Cormier S.M. (2002) Methods development and use of macroinvertebrates as indicators of ecological conditions for streams in the Mid-Atlantic Highlands Region. *Environmental Monitoring and Assessment*, **78**, 169–212.

Klemm D.J., Blocksom K.A., Fulk F.A., Herlihy A.T., Hughes R.M., Kaufmann P.R., Peck D.V., Stoddard J.L., Thoeny W.T. & Griffith M.B. (2003) Development and evaluation of a macroinvertebrate biotic integrity index (MBII) for regionally assessing Mid-Atlantic Highlands Streams. *Environmental Management*, **31**, 656–669.

Linke S., Norris R.H., Faith D.P. & Stockwell D. (2005) ANNA: A new prediction method for bioassessment programs. *Freshwater Biology*, **50**, 147–158.

Montgomery D.C., Peck E.A. & Vining G.G. (2006) *Introduction to Linear Regression Analysis*, 4th edn. John Wiley & Sons, Hoboken, NJ, U.S.A.

Nichols S., Sloane P., Coysh J., Williams C. & Norris R. (2000) *Australian Capital Territory, AUStralian RIVer Assessment System*. Cooperative Research Center for Freshwater Ecology, University of Canberra ACT2601.

Reynoldson T.B., Bailey R.C., Day K.E. & Norris R.H. (1995) Biological guidelines for freshwater sediment based on Benthic Assessment of SedimenT (the BEAST) using a multivariate approach for predicting biological state. *Australian Journal of Ecology*, **20**, 198–219.

Reynoldson T., Smith E.P. & Bailer J. (2002) A comparison of weight of evidence approaches. *Human and Ecological Risk Assessment*, **8**, 1613–1624.

Simpson J. & Norris R.H. (2000) In: Biological assessment of water quality: development of AUSRIVAS models and Outputs. In: *RIVPACS and Similar Techniques for Assessing the Biological Quality of Freshwaters* (Eds J.F. Wright, D.W. Sutcliffe & M.T. Furse), pp. 125–142. Freshwater Biological Association and Environment Agency, Ableside, Cumbria, U.K.

Smith E.P., Robinson T., Field L.J. & Norton S.B. (2003) Predicting sediment toxicity using logistic regression: a concentration addition approach. *Environmental Toxicology and Chemistry*, **22**, 565–575.

Waite I.R., Herlihy A.T., Larsen D.P. & Klemm D.J. (2000) Comparing strengths of geographic and nongeographic classifications of stream benthic macroinvertebrates in the Mid-Atlantic Highlands, USA. *Journal of the North American Benthological Society*, **19**, 429–441.

Wilson D.R. & Martinez T.B. (1997) Improved heterogeneous distance functions. *Journal of Artificial Intelligence Research*, **6**, 1–34.

Wright J.F., Moss D., Armitage P.D. & Furse M.T. (1984) A preliminary classification of running-water sites in Great Britain based on macroinvertebrate species and the prediction of community type using environmental data. *Freshwater Biology*, **14**, 221–256.

Yuan L.L. & Norton S.B. (2003) Comparing responses of macroinvertebrate metrics to increasing stress. *Journal of the North American Benthological Society*, **22**, 308–322.

(Manuscript accepted 10 October 2006)